# ABSTRACT

## AUTOMATIC SELECTION OF MEDIATING ONTOLOGY FOR ALIGNING BIOMEDICAL ONTOLOGIES

by Weiguo Xia

Ontologies are increasingly important in the Semantic Web and biomedical information system fields. Ontology alignment (OA) is the process of finding semantic mappings between the concepts of two given ontologies. OA systems have begun using mediating ontologies pre-selected to improve OA performance. This research investigates the automatic selection of a set of mediating ontologies from a large set of ontologies in the biomedical domain. BioPortal, an online library offering biomedical ontologies via web API and web browsing, is used as the background knowledge source. The anatomy and the large biomedical ontologies tracks of the Ontology Alignment Evaluation Initiative are used to evaluate this approach which is implemented as a software component accessed from a leading OA system LogMap. The experimental results show automatically selected mediating ontologies improve the recall and f-measure for the anatomy track. For the large biomedical ontologies track, three of the six tasks show some overall improvement.

# AUTOMATIC SELECTION OF MEDIATING ONTOLOGY FOR ALIGNING BIOMEDICAL ONTOLOGIES

A Thesis

Submitted to the

Faculty of Miami University

in partial fulfillment of

the requirements for the degree of

Master of Computer Science

Department of Computer Science and Software Engineering

by

Weiguo Xia

Miami University

Oxford, Ohio

2015

Advisor_____

Valerie Cross, Ph.D.

Reader_____

Dhananjai Rao, Ph.D.

Reader_____

Ernesto Jimenez-Ruiz, Ph.D.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my research advisor Dr. Cross for the continuous support of my Master study in Miami University and life in Oxford, OH, for her understanding, patience and immense knowledge. Without her help, I cannot finish this thesis work. Also a very special thanks to Dr. Ernesto who is the creator of LogMap ontology alignment system for his support of using his leading OA system and brilliant idea. I must also acknowledge Dr.Rao for his suggestion on my career and serving as my thesis committee. Last but not the least, I would like to thank my family's support.

# 1 INTRODUCTION

In computer science, ontologies refer to a formal, explicit specification of a shared conceptualization(Gruber 1993). They formally describe domain concepts and their relationships in a machine-readable way and hence have become increasingly important in the Semantic Web and biomedical information systems fields. Ontology alignment (OA) is the process of finding semantic mappings between the concepts of two given ontologies. Many information processing applications such as those in e-commerce, bioinformatics, and knowledge management use multiple ontologies and require establishing these mappings to ensure interoperability. Most OA systems automatically perform this process though some have the option of user interaction in the alignment process.

In the past decade, numerous OA systems have been developed. The Ontology Alignment Evaluation Initiative (OAEI) is an international initiative that provides a systematic evaluation platform for such OA systems (Shvaiko and Euzenat 2013). The aim of the OAEI event is to determine the advantages and drawbacks of OA systems and compare their performances.

The OAEI has different tracks with different ontologies to test the OA systems on a variety of domains and with varying characteristics. In recent OAEI competitions, several of the OA systems have incorporated the use of general background knowledge sources such as WordNet and domain-specific background knowledge such as Uberon or UMLS for the OAEI anatomy track and large biomedical ontologies track (Grau et al. 2013). One of the basic algorithms looks for synonyms in the background knowledge sources for both a source ontology concept and a target ontology concept. If both concepts have a synonym in common, then a mapping between the two concepts is created. In (Silwal 2012), semantic similarity is used within a reference ontology to find alignments even when there is not an exact match on the synonym or concept in the reference or mediating ontology.

In these OA systems, especially for the domain-specific tracks such as anatomy, the selection of ontologies to serve as background or mediating ontologies has been predetermined. The objective of this thesis research is to develop a software component Biomedical Mediating Finder that can be added to an OA system that automatically selects from a set of ontologies the most appropriate ones to use as mediating ontologies to align ontologies in the biomedical domain. Specifically, this component is first added to the LogMap OA system (Jiménez-Ruiz et al. 2012) and BioPortal (Rubin et al. 2008) is provides the set of ontologies from which to select the mediating ontologies.

LogMap was selected as the initial OA system to use Biomedical Mediating Finder with since it had been one of the top performers in the OAEI since 2012. Another factor is that Dr.Ernesto, LogMap's principal architect, has had a close working relationship with Miami University Computer Science Department and his software has already been used for several other completed computer science masters theses.

To further demonstrate its flexibility, the software component developed for this thesis is also used with several other OA system that have participated in the OAEI. The goal is that an OA systems could then use the selected mediating ontologies to improve its alignment process.

This thesis research makes the following contributions:

1. Most of OA system that apply mediating ontology techniques require an expert to pre-select the mediating ontology for specific source and target ontologies being aligned. This thesis research provides a simple and effective algorithm that uses the Restful API of BioPortal to provide a list of suitable mediating ontologies instead of requiring a pre-selected list from an expert.
2. A systematic approach is used that demonstrates the algorithm's implementation to appropriately select the mediating ontologies for the alignment process. The implementation is used first with LogMap and then with other OA systems to show it the software component can be used with other existing OA system.
3. This thesis has used the official OAEI anatomy and large biomedical tracks to verify that the algorithm selects mediating ontologies from BioPortal that produce

significant improvements in the F-measure and recall for the anatomy track of OAEI.

## 2 ONTOLOGY ALIGNMENT TECHNIQUES

Ontology alignment is the process that finds semantic matches between the concepts in different ontologies. These matches can be used for data translation, ontology merging or retrieval of information. Figure 1 illustrates the basic process of ontology alignment. At first, the OA system is given two input ontologies O1 and O2 and an existing set of alignments A, which can be used to find additional alignments. After the alignment process, A' is the resulting alignment. . Some OA systems also receive parameters and external resources. (Taye and Alalwan 2010), for example, background knowledge to assist with the alignment process.

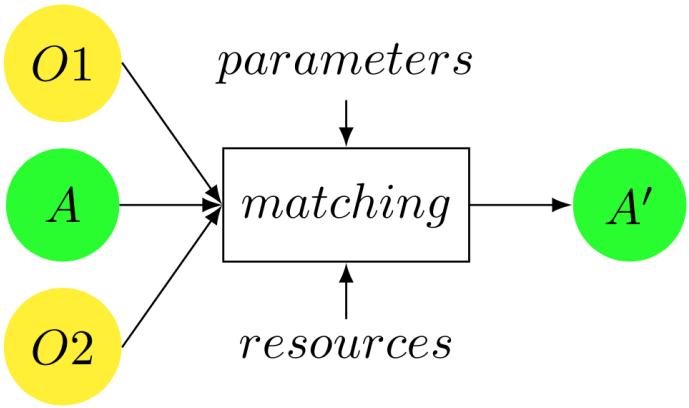


**Figure 1 Ontology Alignment process** (Euzenat and Shvaiko 2007)

Because of the increasing need for ontology alignment, much research has been done in this field in the last decade. . The results from the last several years of the OAEI competition indicate that the performance of OA systems has greatly improved. Much of the improvement can be attributed to improved matcher algorithms and the use of background knowledge resources. The following sections provide a discussion of the

techniques that are used for ontology alignment and the process of evaluating OA systems.

## 2.1 General architecture of OA systems

Matcher algorithms rely on methods to determine the similarity between a source entity and a target entity. Determining similarity occurs at two different levels, the entity itself, i.e., the element level and the structure surrounding the entity.

### 2.1.1 Element-level techniques

In order to find similar entity, the most basic method used is some terminological methods to compare the string label of the entities(Euzenat and Shvaiko 2007). The following are two main element-level techniques:

- String-based methods

  By using the composition of a string, string–based methods usually can find the two strings, for example, football and English football are similar. Before comparing the strings, the strings usually are normalized in order to improve the accuracy of the matching, for example, case normalization and removing multiple blank characters. There are numerous string-based matching methods. Two standard ones are Hamming distance and substring similarity. Hamming distance: is a mathematical way that compares the two strings by counting the number of different character. Substring similarity for any two string x and y is given as,

  $$s = \frac{2|t|}{|x|+|y|}$$

  which t is longest common substring of x and y.

- Language-based methods

  In addition to the composition of the string label for an entity, there is a grammatical structure. Therefore, matchers can take advantage of exacting useful text by using Natural Language processing (NLP) techniques. The following are two main methods:

*Intrinsic methods:* Linguistic normalization is process that standardizes the string. Linguistic software is used and typically uses the following steps: parse, tokenization, lemmatization, term extraction and stop word elimination.

*Extrinsic methods:* Extrinsic methods use external linguistic resources to find similarities between terms such as lexicons, muti-lingual lexicons, semantico-syntacitic lexicons, thesauri and terminologies. The main purpose of introducing the external linguistic resources is to identify the synonyms.

Terminological methods are usually very effective; however, synonyms and homonyms present difficulties when comparing the labels used for entities in the ontologies to be aligned. Most OA systems, therefore, apply some structure-level techniques.

### 2.1.2 Structure-level techniques

Beside the comparison of individual terms, the structures of the term in the target and source ontology are also comparing to find mappings.

- Internal structure

The internal structure means comparison based on its term name, properties and annotations without using information of other entities. The similarity between two terms is calculated based on the set of their properties, such as data type of the property.

**Table 1 Part of a datatype compatibility table.**(Euzenat and Shvaiko 2007)

| | Char | Fixed | Enumeration | Int | Number | string |
|---|---|---|---|---|---|---|
| String | 0.7 | 0.4 | 0.7 | 0.4 | 0.5 | 1.0 |
| Number | 0.6 | 0.9 | 0.0 | 0.9 | 1.0 | 0.5 |

This technique is mainly used on database matching. But in ontology matching, this technique doesn't provide much information. Because many different terms in ontology with same datatype due to restrictions in OWL, the ontology definition language.

- Relational structure

Ontology can be seen as a directed graph with the edges representing the relationships between the entities, which are represented by the nodes in the graph. For the relational structure approach, matching two ontologies use techniques such as finding the maximum common graph of two-target graph representing the ontologies.

## 2.2 Overview of BACKGROUND KNOWLEDGE

Although research has initially focused on element-level and structure-level techniques that have been shown to be effective for ontology alignment, the need to improve the performance of OA systems caused research to turn to the use of background knowledge sources. The following sections describe several of the background knowledge sources used in recent OAEI competitions.

### 2.2.1 Wikipedia

Wikipedia is a free content encyclopedia which is openly editable for everyone through Internet. Presently, Wikipedia contains 4.745,309 articles written by anonymous volunteers in multi-lingual. Its content can also be obtained by using the API service provide by MediaWiki over HTTP. (Nakayama, Hara, and Nishio 2008)

Wikipedia has two main components, articles and categories, each with a different role in Wikipedia. An article contains the main information that most readers are looking at. This information can used by string-based methods of OA systems. A category is used as a classification method to help a user find the target concept from one related article in a related subfield. Wikipedia also maintains a category tree which is a hierarchical dynamic organization that structures the relationships between categories.

Both BLOOMS+(Prateek Jain, Peter Z. Yeh, Kunal Verma, Reymonrod G. Vasquez, Mariana Damova, Pascal Hitzler 2011) and WikiMatch(Hertling and Paulheim 2012) are using Wikipedia as an external knowledge source in the task of aligning linked open data ontologies. In both OA systems Wikipedia's categories are used in finding matches

between concepts in two different ontologies. Whereas, BLOOMS+ uses the category structures of Wikipedia, WikiMatch only looks up concepts in Wikipedia by using Wikipedia's documents. Concepts are described by the normalized string of their fragments, labels or comments. A search query in Wikipedia returns a set of document ids. To determine whether to match two concepts, a set similarity measure is used between their returned document id sets. If the set similarity measure meets a threshold, then the two concepts are determined to be a match. Due to the simpler algorithm of WikiMatch, it use less runtime than BLOOMS+.

### 2.2.2 WordNet

WordNet is a machine-readable English lexical database. It is composed of synsets, which contain a group of synonyms. A example of synset for mess is { fix, hole, jam, muddle, pickle, kettle of fish}. A lexical concept can have several senses; therefore, these words are interchangeable in a sentence in that particular one sense. But these words are not interchangeable in any other sense of the lexical concept. Synsets are connected by different kinds of relationships, such as hypernyms and hyponyms, part-of, derivatives and so on.
WordNet can be accessed over HTTP or installed on a personal computer through command line(Kamps and Safe 1987).

Many researchers have use WordNet as background knowledge when aligning ontologies(Reynaud and Safar 2007) .Usually there are three main way: 1)find the synonyms for a concept of the ontologies being aligned, 2) measure the similarity of two concept, and 3) infer a mapping between two concepts in the ontologies being aligned based on both being equivalent to the same concept within Wordnet.

### 2.2.3 Uberon

In bioinformatics field, anatomy ontologies have been proven usefully for database and bioinformatics analyses. These representations enable automatically inferring

information. But inferred information between different anatomy ontologies is still problematic because the different anatomy ontologies use differing vocabulary and structure. To address this issue, Uberon (Mungall et al. 2012) was created as multi-species anatomy ontology with different versions for different purposes.

The main version of the ontology consists of 6,500 classes representing a set of high-level concepts for all anatomy ontologies. For example, "nervous system" and "circulatory system" are high level concepts for all anatomical systems. The main version uses constructors from OWL2-DL language that enable richer axiomatic knowledge representation. Uberon/ext is the extend version of the main version. It also contain the subset that are from other ontologies such as Cell Ontology (CL) and the Gene Ontology (GO).

Researchers recently are using Uberon as a bridge between different anatomy ontologies. For example, one cannot query for all pharyngeal relationships by using FMA (Foundational Model of Anatomy) or MA(Mouse adult gross anatomy ontology) alone, but with Uberon more of these relationships can be found. This use of Uberon suggested that it could serve as a background knowledge source for ontology alignment.
Several OA systems have been using Uberon ontology, specifically as a mediating ontology in the OAEI anatomy track to align the Mouse Anatomy ontology and Human Anatomy ontology(José-Luis Aguirre, Kai Eckert, Jérôme Euzenat, Alfio Ferrara, Willem Robert van Hage, Laura Hollink, Christian Meilicke, Andriy Nikolov, Dominique Ritze, François Scharffe, Pavel Shvaiko, Ondrej Sváb-Zamazal, Cássia Trojahn dos Santos, Ernesto Jiménez-Rui 2012). AgreementMaker(Cruz, I. F., Stroe C., Caimi F., Fabiani A., Pesquita C., Couto F. 2011) has also used Uberon as a mediating ontology in its lexicon framework.

### 2.3 Evaluation

In the information retrieval (IR) field, precision, recall and the F-measure (Euzenat 2007) are common measures used to judge the performance of an IR system. These standard

measures have been adapted for ontology alignment and are based on a reference set of alignments, also referred to as a gold standard. The reference alignment R is created by experts for the specific source and target ontologies being aligned. The reference alignment is regarded as the correct alignment result for the source and target ontologies. In addition to these standard performance measures, the computation time taken by the OA system and its coherence are determined. Coherence is based on the number of satisfiable classes as determined using a reasoning system on the input ontologies and the mappings produced by the OA system.

Given the reference alignment R and an alignment A produced by an OA system, precision is defined as:

$$P(A,R) = \frac{|R \cap A|}{|A|}$$

It is the fraction of mappings in the produced alignment A that are considered correct based on the reference alignment R.

Recall is defined similarly but the denominator changes to the number of mappings specified in the reference alignment R:

$$R(A,R) = \frac{|R \cap A|}{|R|}$$

It is the fraction of the mappings in the reference alignment R that are produced in the alignment A. The F-measure combines precision and recall using a parameter $\beta$:

$$F_\beta = (1+\beta^2) \cdot \frac{P(A,R) \cdot R(A,R)}{\beta^2 P(A,R) + R(A,R)}$$

The weighting factor β allows emphasizing either recall or precision. When β is 1, $F_1$ gives equal importance to precision and recall and the result is the harmonic mean of the two

## 2.4 State of the art OA systems: LogMap

Numerous OA systems have demonstrated good performance in the OAEI competitions. LogMap has been one of the leading OA systems for the last several years. Its design emphasizes important features such as scalability, inconsistency repair, and

interactivity (Jiménez-Ruiz et al. 2012). LogMap uses an inverted index which stores the lexical information contained in the input ontologies to efficiently match semantically rich and large ontologies containing up to hundreds of thousands of classes. LogMap determines an initial set of mappings of manageable size using this index. Sophisticated reasoning and repair techniques are used to minimize the number of logical inconsistencies produced by equivalence mapping in the alignment. For some mappings with lower confidence an interactive step may be used to allow the expert user to provide input to the alignment process.

LogMap has two primary phases. The objective of the first phase is to optimize recall and for the second phase to optimize precision. This first phase consists of four steps: lexical indexation for each input ontology, computation of candidate class mappings using the lexical indexes, computation of candidate property mappings by performing a pairwise string comparison (with ISUB , a substring similarity method) on the URIs and labels of properties from the two ontologies, and logic-based module extraction to produce smaller fragments of the input ontologies using the candidate mappings. The lexical indexes for the source and target ontologies are intersected to find the initial set of equivalence candidates. LogMap considers these mappings as accurate and uses them as a start for finding more mappings. Extracting smaller modules by using the candidate mappings improves the efficiency of the unsatisfiability detection and repair algorithms.

The second phase is more complex than the first. Not all candidate mappings found in the first phase are reliable. To be reliable, a mapping must have a high confidence level which is determined by a string matcher used on the source class and the target class. In addition, at least one child (parent) of the source class must map to at least one child (parent) of the target class. Although a candidate mapping may be found to be reliable, if a few incorrect reliable mappings exists, they can cause many classes to become unsatisfiable. LogMap detects unsatisfiable classes by encoding the extracted smaller modules and the reliable mappings into Horn propositional clauses and then uses an extended Dowling-Gallier algorithm (Dowling and Gallier 1984) for propositional Horn satisfiability. LogMap's extension to this algorithm records the conflictive mappings. The conflictive mappings involved in an unsatisfiable class are examined by LogMap in increasing size of the number of conflictive mappings. LogMap tests to see if by

removing them, the class becomes satisfiable. The final result is a set of repaired reliable mappings.

The non-reliable mappings are then processed efficiently by using the lexical index and the semantic index for the classes in the extracted modules and in the repaired reliable mappings. If a non-reliable mapping is added to the reliable mappings and causes a class to become unsatisfiable, the non-reliable mapping is removed. Other algorithms are also used to remove non-reliable mappings. For the remaining non-reliable mappings, their confidence values may be revised using both the lexical index and the semantic index. Co-occurrence analysis using the lexical index on terms in the class names in the non-reliable mappings is performed. A high co-occurrence causes the confidence of that mapping to be increased. The principle of locality states that if two classes are mapped to each other, then their superclasses and subclasses are likely to be mapped. The confidence of the mapping is increased if the principle of locality holds for a non-reliable mapping.

To reduce the number of the non-reliable mappings, LogMap may enter a user interaction step. Using the revised confidence values for the non-reliable mappings, a partial order is created which is used to present questions to the user. The user may accept or reject a given mapping. The user can also decide to stop the interactive process and allow LogMap to heuristically decide the remaining mappings. At the end of this interactive step is a set of mappings accepted from the non-reliable mappings and used by LogMap to produce the final mapping results. If user interaction occurs, mappings selected by the users are given priority. This user priority may cause some automatically computed reliable mappings to be deleted. When there is no user interaction, reliable mappings may not be deleted and, thus, take precedence over the remaining non-reliable mappings.

Currently, LogMap uses only one kind of external knowledge source, the UMLS Lexicon, which is a lexicon that offers a set of variants for each lexicon entry. Logmap uses this lexicon since it includes a diverse collection of biomedical words and enriches Logmap's index by finding more spelling variants.

# 3 RELATED RESEARCH

Ontology alignment can benefit from background ontologies since some semantic relationships may be found that are not recognize otherwise. Several OA systems participating in recent OAEI competitions have used one or more background knowledge sources. There also has been some research in the automatic selection of background knowledge sources. The following sections discuss these two recent areas of research in ontology alignment.

## 3.1 Using Mediating ontologies

In (Gross et al. 2011), the reference ontology is called an intermediate ontology and in (Cruz, I. F., Stroe C., Caimi F., Fabiani A., Pesquita C., Couto F. 2011) it is referred to as a mediating ontology. Both follow a very similar approach. The only differences exist in the alignment methods used to produce the mappings from the source and target ontologies to the intermediate ontology and what aggregation method of similarity values is used to produce the final mapping from a source concept to a target concept through a concept in the mediating ontology.

The composition-based matching in (Gross et al. 2011) was the first to introduce this approach. In effect, two simplified ontology alignments are first performed to create the mappings between the source ontology and the intermediate ontology and between the target ontology and the intermediate ontology. If a source to intermediate mapping concept matches a target to intermediate mapping concept then a mapping between the source and target concepts is added to the alignment results. The composition approach was evaluated using the OAEI anatomy track and four separate and combined intermediate ontologies: FMA, Uberon, RadLex, and UMLS. From their experiments Uberon produced the best results.

AgreementMaker was then modified to follow the lead of the composition-based matching approach. A mediating matcher (MM) was added to the system (Cruz, I. F., Stroe C., Caimi F., Fabiani A., Pesquita C., Couto F. 2011) for its participation in OAEI 2011. This OA system uses a variety of matching algorithms and integrates the results of these individual matchers and hence can perform well in different scenarios. Based on

Uberon having the best results previously (Gross et al. 2011), the MM used Uberon as its mediating ontology to expand its lexicon for the OAEI anatomy track. Another reason for the selection of Uberon is that it is a cross species anatomy ontology. Its use increased AgreementMaker's precision by over 5%.

Figure 2 illustrates the general steps using a mediating or intermediate ontology $O_I$ to assist in aligning the source ontology $O_S$ and target ontology $O_T$.

(1) Use a simple and quick matcher to produce a set of mapping $M_{SI}$ between $O_S$ and $O_I$

(2) Use a simple and quick matcher to produce a set of mapping $M_{TI}$ between $O_T$ and $O_I$

(3) When a source concept *s* and a target concept *t* map to exactly the same bridge concept in the mediating ontology $b_s = b_t$, add an equivalence mapping between *s* and *t* to the set of output mappings $M_{ST}$ produced by the MM

(4) Integrate the mappings $M_{ST}$ with the mappings produced by the other matchers in AgreementMaker

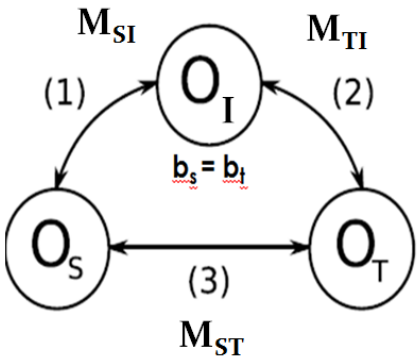


**Figure 2 Using Mediating Ontology**(Cruz, I. F., Stroe C., Caimi F., Fabiani A., Pesquita C., Couto F. 2011)

In (Silwal 2012), AgreementMaker's original mediating matcher was replaced by the mediating matcher with semantic similarity measure (MMSS). MMSS executes the mediating matcher first and saves the result mappings. Then it check all the source concepts mapped to mediating ontology and target concept mapped to mediating

ontology that do not have an exact match. It uses a semantic similarity measure between the mediating ontology concepts to see if two of them are close enough to add an additional mapping between a source concept and a target concept. The ontologies FMA and Uberon were pre-selected as the mediating ontologies. The objective of the thesis research proposed here is to automatically determine the most appropriate ontologies to use as mediating ontologies.

## 3.2 Automatic Selecting Background

The previous research briefly described in the previous section has shown that using a mediating ontology can improve the results of ontology alignment. The mediating ontologies are typically in the same domain as the source and target ontologies and have been pre-selected specifically for the ontology alignment task. Research (Quix, Roy, and Kensche 2011) has investigated automatically selecting appropriate ontologies as background knowledge. Fig. 3 illustrates the overall approach implemented in an existing matching system GeRoMeSuite (Kensche et al. 2007).

For an input source ontology S and target ontology T, two queries $Q_S$ and $Q_T$ are developed to represent the source and target ontologies, respectively. The vector space (VS) information retrieval model approach is used to check whether a document looks similar to another document in an efficient and scalable way. Each ontology in the local repository is translated to a corresponding background document (BGdoc). A BGdoc is created by extracting information such as concept names, comments and labels from the ontology and applying some text processing, like stemming and tokenization. The Apache Lucene (Addagada 2007) is used to apply the VS retrieval model for $Q_S$ and $Q_T$ to search the local ontology repository. A ranked list of ontologies in the repository is returned with an information retrieval similarity score. The ontologies that maximize the formula

$$\alpha(\text{sim}(O,S)+\text{sim}(O,T))-\beta|\text{sim}(O,S)-\text{sim}(O,T)|$$

are then selected as background knowledge sources for the alignment. This formula can select an ontology that,

(1) is most similar to S and T (i.e., maximize sim(O; S) + sim(O; T)),

(2) is similar to both ontologies, and not only to one (i.e., minimize |sim(O; S) - sim(O; T)|), and

(3) should meet a minimal similarity threshold of the ontology to both S and T, i.e., thresholds for sim(O, S) and sim(O, T).

A high value for α does not perform well since it prefers ontologies that are very similar to one of the input ontologies and highly dissimilar to the other. Hence, the α is set to a value only slightly higher than β. Experiments were used to set the threshold for the required minimal similarity

If the local repository does not have appropriate ontologies, a single query is generated to represent the source and target ontologies and used to query the Web. The Web external search engines just return a ranked list of results without similarity scores. Standard web search engines are used and the queries are specifically used on ontologies represented using OWL and RDF. Google and also other ontology search engines such as Swoogle (Li Ding, Rong Pan, Tim Finin, Anupam Joshi, Yun Peng 2005) and Watson (Quix, Roy, and Kensche 2011) were used. Experiments showed that the top results returned by these systems often are very general ontologies and are not useful as background knowledge sources in ontology alignment.

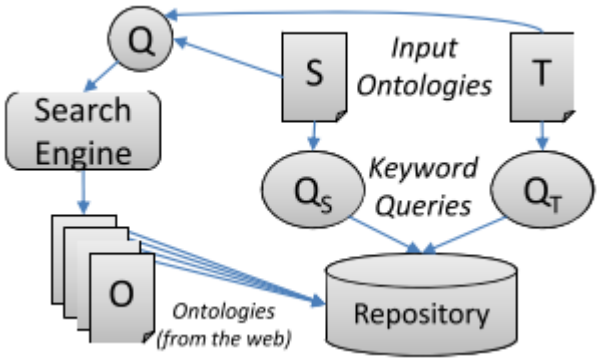


**Figure 3 Selection of the background ontology**(Quix, Roy, and Kensche 2011)

# 4 BIOPORTAL AS BACKGROUND KNOWLEDGE

The research in (Kensche et al. 2007) used a repository of ontologies to select from for the role of mediating ontologies and then used the Web to search for one if those in the repository were not satisfactory. The approach taken in this thesis research is similar in that a repository is searched to find appropriate mediating ontologies to improve the ontology alignment process. An emphasis in OAEI competitions and recent research has been in aligning biomedical ontologies; therefore, a suitable repository for this domain is BioPortal.

## 4.1 Overview of BioPortal

BioPortal is an online library that offers biomedical ontologies via a Web API and Web browsing (Noy et al. 2009). The biomedical ontologies in BioPortal are in RDF, OWL, or OBO format. BioPortal offers its users browsing, searching and visualization of ontologies.

BioPortal has been selected as the repository from which to select mediating ontologies for the biomedical domain because it contains more than 360 ontologies that cover many different areas within this domain such as anatomy, phenotype, experimental conditions, imaging, chemistry, and health. This variety means that a wide range of source and target biomedical ontologies may be input to LogMap and use the BioPortal repository. BioPortal includes the Uberon ontology, for example. This fact is used to perform experiments were with the OAEI anatomy track to show the selection of Uberon and other anatomy ontologies from BioPortal.

BioPortal offers an open REST service API which can be called to access the ontologies and their metadata. This interface is used to query BioPortal for information needed for automatically selecting the mediating ontologies to use in the alignment process.

## 4.2 An Algorithm for Select Mediating Ontologies

This research implements a software component that can be added to LogMap and other OA systems. Its purpose is to automatically select ontologies as mediating ontologies for aligning source ontology and target ontology. BioPortal serves as a fixed repository from which to select the mediating ontologies.

This research differs from that in (Quix, Roy, and Kensche 2011). It uses efficient querying of the metadata which describes BioPortal ontologies. The complete source ontology and target ontology are not represented as document descriptions, but instead the reliable candidate mappings and the smaller extracted ontology modules from the source and target ontologies are used to determine sets of concepts that can be used to query BioPortal to find the candidate mediating ontologies. Note that exact mappings M is same as the initial mappings between the source and target ontology that LogMap has produced based on the source concept and target concept having high lexical similarity. These are also referred to as reliable equivalence mapping.

The following algorithm is used:

Algorithm 1 Algorithm to identify mediating ontologies from BIOPORTAL
Input: O1, O2: input ontologies; LM: a lexical matcher; N: stop condition
Output: Top-5 (candidate) mediating ontologies MO
1: Compute set of exact mappings M between O1 and O2 using a lexical matcher LM. These take the form <s, t, equiv>
2: For each $<s_i, t_i, equiv_i>$ in M
3: If i is even
4: then add $s_i$ to set Se
5: else add $t_i$ to set Se
6: For each label *l* in set Se
7: Get ontologies from BIOPORTAL that contains an entity with label *l* (search call)
8: If the ontology does not already exist in MO
9. then add to MO,
10: record 1 positive hit count,
11: record number of synonyms
12: record ontology information: # of classes, depth and DL expressiveness
13: else increment by 1 positive hit count
14: increment # of synonyms by # of synonyms returned by search call
15: Reorder list of ontologies based on # positive hits and # of synonyms
11: Loop stop condition:
if after N calls to BIOPORTAL, the top 5 MO do not change then stop iteration

12: return Top-5 ontologies from MO according to the number of positive hits

The alternating on even and odd for using the label from the source or target in the exact mappings M is to try to make the mediating ontology relate as much as possible to both the source and the target ontology. Adding the labels from only the source ontology might cause the mediating ontology to have more semantically similarity to the source ontology than to the target ontology. This result can damage the ability of finding mappings using selected mediating ontology.

## 4.3 Initial Experiments with LogMap on anatomy track

The software implementation of the algorithm was run on the anatomy track as a software component added to LogMap and using BioPortal as the mediating ontology repository.

The list of the top five mediating ontologies resulting from this initial experiment is shown in Table 2. This list is ranked based on the number of concept queries a mediating ontology successfully responded to when queried and the average number of synonyms that it provides over all the successful concept queries made to BioPortal by the software component added to LogMap. This data is shown in Table 3. For this initial experiment, it was believed that using the number of synonyms to determine the better mediating ontologies would be useful but as more experiments were performed the percentage of concepts found in the mediating ontology was a better and more useful criteria for ranking the mediating ontologies. In this section, however, the initial experiment is discussed.

The Total Mappings column specifies the total number of mappings produced by LogMap when using the mediating ontology given in the first column. The column New Mappings from Mediating Ontology (MO) specifies the number of mappings produced by LogMap using the mediating ontology and not found by the original LogMap. The column correct mapping indicates the number of mappings that are correct in the set of new mappings. As shown in the table, Uberon and SNOMED CT indeed find many good

mappings that were not found in the original results of LogMap. However, the mediating ontologies produce new mappings that are not correct. If all the new mappings found are included in the final alignment results, the recall will increase but the precision will decrease. The results of this initial experiment indicated the need for more research to discover how to balance the recall and precision results to produce a better F-measure.

**Table 2 The Top 5 Mediating ontologies (MO) that current algorithm find and the new mapping produced by LogMap using them.**

| MO | Total Mappings | New Mappings from MO | Correct mappings in New Mappings |
|---|---|---|---|
| SNOMED CT | 1589 | 321 | 151 |
| Uberon | 603 | 133 | 38 |
| MESH | 236 | 5 | 0 |
| EFO | 238 | 29 | 2 |
| CL | 126 | 26 | 4 |

**Table 3 Top 5 mediating (BIOPORTAL) ontologies (MO) for the OAEI's anatomy track**

| MO | % found concepts in MO | Avg. # syn. | # classes |
|---|---|---|---|
| SNOMED CT | 60% | 5.1 | 40122 |
| Uberon | 63% | 3.3 | 12091 |
| MESH | 34% | 5.0 | 232262 |
| EFO | 16% | 5.1 | 14253 |
| CL | 22% | 3.3 | 5534 |

Table 3 shows the criteria used to rank the top ontologies which are the percentage of concepts from the ontologies being aligned that were found in the MO (% Found concepts in the MO column) and the average number of synonyms existing in the MO for

each found concept (Avg. # syn. column). For example, the Uberon mediating ontology produced at least one synonym for a concept query in 63% of the queries. On average Uberon produces 3.3 synonyms for each of its successful concept queries. The column number of classes show how many classes the mediating ontology contains.

As seen in the table SNOWMED CT and Uberon on very close in the % Found column but SNOWMED CT produces more synonyms. Similarly, CL had a higher % Found than EFO but EFO produces more synonyms. This ranking of the mediating ontologies is a rough estimate using those two criteria. Through more experiments and the actual results of the alignment process the % Found criterion proves more useful.

These experiments demonstrate that a mediating ontology can assist an OA system in finding new and correct mappings not found by its own set of matchers. More experiments are described in the next section to determine how to best (increase recall but decrease precision as little as possible) incorporate these new mappings from the BioPortal mediating ontologies into the results of an OA system.

# 5 Evaluation with LogMap And BioPortal

Since this thesis research focuses on using BioPortal (Rubin et al. 2008) as background knowledge for the biomedical domain, the anatomy track and the large biomedical ontologies track are the relevant OAEI tracks to evaluate software component added to an OA system to automatically select mediating ontologies. The anatomy track has the objective of aligning the Mouse Anatomy with the NCI Thesaurus ontology that describes human anatomy. In the large biomedical ontologies track, OA systems are to find mappings between the Foundational Model of Anatomy (FMA), SNOMED CT, and the National Cancer Institute Thesaurus (NCI). These three ontologies are semantically rich and have tens of thousands of concepts so that the OA systems are evaluated on real world sized ontologies.

## 5.1 Anatomy Track

One feature of this track is a high number of trivial mapping that can be found just by using simple string matching methods. In contrast, there are also some non-trivial mapping that require deeper mining and additional biomedical background knowledge (Dragisic et al. 2014). Therefore, the use of mediating ontologies should produce improved performance by OA systems for this track. Table 4 shows that this is the case. It compares the precision, recall and F-measure of the original LogMap and LogMap using the selected mediating ontologies from BioPortal. LogMap with mediating ontologies from BioPortal include all the mapping from original LogMap results and integrates some of the other mappings produced by the mediating ontologies. The final results shown in Table 4 are produced after several experiments which are discussed below in order to better tune the automatic selection of mediating ontologies and to incorporate the use of more than one mediating ontology.

**Table 4: Original LogMap vs. LogMap using the BioPortal respository on the OAEI Anatomy Track**

| | PRECISION | RECALL | F-MEASURE |
|---|---|---|---|
| LogMap | 0.9125 | 0.8463 | 0.8782 |
| LogMap-Bio | 0.8793 | 0.9037 | 0.8913 |

As often occurs when increasing the recall, the precision decreases because finding more of the correct mappings is a result of finding more mappings in general, some of which are incorrect ones. In the following, the approach taken to using BioPortal as a source of mediating ontologies for the anatomy track is described. These experiments and analysis contributed to the improvement in the recall and overall f-measure seen in Table 4.

Table 5 shows the top 5 mediating ontology ranked based on the number of concept queries a mediating ontology successfully responded to when queried, that is, positive hits shown in column 2. Positive hits are used to rank the mediating ontology since it reflects the lexical overlapping between the mediating ontology and the target and source ontologies. Because all concepts that are used to query BioPortal come from the reliable set of mappings calculated by LogMap, concepts in this set are very likely correct

mappings between the target and source ontology. For example, the reliable mappings produced by LogMap for FMA-NCI had a 0.90 precision (Jiménez-Ruiz et al. 2012).

A positive hit means that the concept from the reliable set of mappings calculated by LogMap and used to query BioPortal can be found in the mediating ontology. Therefore, the mediating ontology can more likely provide synonyms that can be used to assist when trying to align concepts between the target and source ontologies.

The MA (Mouse Adult Gross Anatomy) ontology responded 133 times when queried for concepts that are in the set of the original equivalence concepts of the reliable mappings produced by LogMap. Based on the synonyms returned by a mediating ontology, the OA process is able to produce additional mappings between the source and target ontologies. Table 5 calculates the performance measures based on only the "composed" mappings produced by using the mediating ontology. The following algorithm is used to produce composed mappings through the mediating ontology.

Algorithm 2 Algorithm to produce composed mappings by using mediating ontology
Input: O1, O2: input ontologies; OA: OA systems; M: mediating ontology
Output: a set of mappings M3
c1 is the confidence degree for a mapping between the source and mediating ontologies,
c2 is the confidence degree for a mapping between the target and mediating ontologies,

1: Compute mapping set M1 (m, s, c1) by run OA(M,O1)
2: Compute mapping set M2 (m, t, c2) by run OA(M,O2)
3: for $map_1$ (m1, s, $c_1$) in M1:
4: for $map_2$ (m2, t, $c_2$) in M2:
5: if m1= m2: add map(s, t, $(c_1+c_2)/2$ ) to composed mapping M3
6: return M3

The reason the top 5 mediating ontology in table 2 and table 5 differ is some ontologies cannot be download from BioPortal, such as SNOMED CT. In section 4 only the algorithm for finding top mediating ontologies was examined. In this section the results produced are based on actually being able to use the mediating ontology. Also, in Table 2 the

ontologies were ranked by the average of positive hits and the total number of synonyms. In Table 5 only positive hits were used to rank the mediating ontology.

**Table 5 Top 5 mediating ontology for anatomy track**

| | Positive hits | Composed mappings from Mediating ontology | | |
|---|---|---|---|---|
| | | PRECISION | RECALL | F-MEASURE |
| MA | 133 | 0.9240 | 0.8100 | 0.8633 |
| SYN | 122 | 0.8847 | 0.8456 | 0.8648 |
| UBERON | 77 | 0.8571 | 0.9024 | 0.8792 |
| CL | 27 | 0.7770 | 0.2091 | 0.3295 |
| EHDAA2 | 27 | 0.9555 | 0.1557 | 0.2677 |

The composed mappings from mediating ontology are just the mappings that can be found by using the mediating ontology. That is why the recall is much lower than the recall result from the original LogMap. The composed mappings also contain some overlap with the result of the original LogMap which does not use any mediating ontologies.

The positive hits count of Uberon is lower than that of the MA and SYN mediating ontologies yet the performance of Uberon is the best in Table 5. Although MA and SYN seems to have more intersection in terms of concepts with the source and target ontologies, Uberon is providing a richer set of synonyms which is very helpful to finding composed mappings using Uberon as a mediating ontology between MA and NCI.

The results in Table 5 are based on using each of those mediating ontologies separately and show the algorithm can find qualified mediating ontologies that can add correct new mappings. A challenge is how to include correct mappings and minimize the addition of incorrect mappings in the set of all new mappings found by using a mediating ontology. That is, the objective is to keep precision from decreasing too much with incorrect mappings while trying to increase recall.

To determine a possible answer, the results of another experiment are shown in Table 6 and Table 7. The rationale for this test case is the more mediating ontologies that produce a new mapping, the more likely that new mapping is to be correct. The objective is to try to keep only the correct mappings from all the new mappings produced by the mediating ontologies so that recall can be increased and not decrease precision, thus, increasing the f-measure. The process is to examine collectively the suggested new mappings from all the selected mediating ontologies.

The rows in the table indicate the minimum number of mediating ontologies that found the new mapping. Row 1 indicates that a new mapping is added if at least one mediating ontology produced the new mapping. Row 2 indicates that a new mapping is added if at least two mediating ontologies produced the new mapping and so on. The column precision, recall and F-Measure indicate the performance of requiring that minimum number of mediating ontologies to produce the new mapping. As show in Table 6 when the minimum number is 2, the f-measure performance is the greatest so that in the following testing, at least two mediating ontologies must produce the new mapping for it to be added to the final mapping results.

**Table 6 Test case to choose minimum number.**

| | Precision | Recall | F-Measure |
|---|---|---|---|
| 1 | 0.7899 | 0.9472 | 0.8614 |
| 2 | 0.8940 | 0.8734 | 0.8836 |
| 3 | 0.9095 | 0.8489 | 0.8782 |
| 4 | 0.9112 | 0.8463 | 0.8776 |
| 5 | 0.9125 | 0.8463 | 0.8782 |

As can be seen from table 6, precision was increased greatly from row 1 to row 2, but recall decreased significantly. The next experiment is undertaken to find correct new mappings without eliminating those produced by only one mediating ontology since recall is highest when only one mediating ontology is needed to include a mapping. To

accomplish this objective, another condition, the confidence degree of the mapping is checked to decide if a new mapping should be considered a good mapping.

In Table 7 row 1 shows the results of adding a new mapping if its confidence degree is bigger than 0.7 for one mediating ontology or at least two mediating ontologies produce the mapping. Row 2 is similar to row 1 but requires a 0.8 confidence degree if only one mediating ontology produces the mapping. Row 3 is similar to row 2 but requires a 0.7 confidence degree even if at least two mediating ontologies produce the mapping.

As shown, row 3 has a slightly better performance than row 2. This condition is applied to the use of BioPortal mediating ontologies with LogMap to produce the results given in previous Table 4. The reasons that 0.7 was selected as the minimum confidence degree is the precision of the composed mappings found by mediating ontology increased from 0.4322 to 0.7215 when the confidence degree was changed from 0.6 to 0.7. The parameter 0.8 is chosen because the recall of the composed mappings found by mediating ontology has dropped from 0.0375 to 0.00461 but the precision increased from 0.7215 to 0.875. The performance of shown in Table 7 is calculated based on the integration of the composed mappings with the original LogMap alignment results.

**Table 7 Conditions for Filtering Mediating Ontology Mappings**

| | | Precision | Recall | F-Measure |
|---|---|---|---|---|
| 1 | 2 MOs OR (1 MO AND Cf > 0.7) | 0.8420 | 0.9354 | 0.8863 |
| 2 | 2 MOs OR (1 MO AND Cf > 0.8) | 0.8717 | 0.9096 | 0.8903 |
| 3 | (2 MOs AND Cf >0.7) OR (1 MO AND Cf > 0.8) | 0.8793 | 0.9037 | 0.8913 |

## 5.2 Large Biomed Track

As explained previously, there are three ontologies: the Foundational Model of Anatomy (FMA), SNOMED CT, and the National Cancer Institute Thesaurus (NCI). There are two versions of each pair of ontologies, the complete ontology itself and a smaller version of each of the three ontologies referred to as an ontology fragment. Each complete ontology is aligned to the other two complete ontologies, and each smaller ontology version is aligned with the other two smaller ontology versions to produce six alignment tasks in this track.

For each task, there is a table below that records the Top 5 mediating ontologies. These are ranked by the positive hits. The performance measures shown are based on only using the mappings produced by the mediating mappings, i.e., the composed mappings. For example, in Table 8, for small FMA-NCI task, the SYN ontology has 111 positive hits. Its performance result is based on mappings found by using SYN as the mediating ontology.

For each of the six tasks, there is also a corresponding a-version table. For example, for the small FMA-NCI task, Table 8a shows the result of LogMap with and without the selected mediating ontologies. In the a-versions of the tables, the results are based on combining the composed mappings with those mappings produced by LogMap. The condition listed in row 3 of Table 7 is used to filter the mappings produced by the mediating ontologies. In all the tables showing LogMap-Bio results, the precisions are decreasing and the recalls are increasing. For FMA-SNOMED small fragment, SNONMED NCI whole ontologies and SNONMED NCI small fragments the F-measures increased.

In the tables comparing the top 5 mediating ontologies for each of the six large biomedical ontologies tasks, the performance measures are calculated based on only the composed mappings produced by using mediating ontology; therefore, the recall is much lower than the recall found in the “a” versions of the corresponding tables which use both the composed mappings combined with the mappings produced from LogMap.

**Table 8 Top 5 mediating ontology for small FMA-NCI**

| | Positive hits | Composed mappings from Mediating ontology | | |
|---|---|---|---|---|
| | | PRECISION | RECALL | F-MEASURE |
| SYN | 111 | 0.9339 | 0.7986 | 0.8610 |
| UBERON | 57 | 0.9016 | 0.6273 | 0.7399 |
| MA | 44 | 0.9435 | 0.3588 | 0.5199 |
| CL | 36 | 0.8544 | 0.2097 | 0.3367 |
| XAO | 21 | 0.8575 | 0.1174 | 0.2065 |

**Table 8a Small FMA-NCI LogMap to LogMap-Bio**

| | PRECISION | RECALL | F-MEASURE |
|---|---|---|---|
| LogMap | 0.9541 | 0.8598 | 0.9045 |
| LogMap-Bio | 0.9281 | 0.8793 | 0.9030 |

**Table 9 Top 5 mediating ontology for the whole FMA and NCI**

| | Positive hits | Composed mappings from Mediating ontology | | |
|---|---|---|---|---|
| | | PRECISION | RECALL | F-MEASURE |
| SYN | 100 | 0.7161 | 0.7868 | 0.7498 |
| UBERON | 59 | 0.6118 | 0.6104 | 0.6111 |
| MA | 46 | 0.6610 | 0.3480 | 0.4560 |
| CL | 45 | 0.7579 | 0.1890 | 0.3026 |
| XAO | 20 | 0.7674 | 0.1126 | 0.1964 |

**Table 9a Whole FMA –NCI LogMap to LogMap-Bio**

| | PRECISIO N | RECALL | F-MEASURE |
|---|---|---|---|
| LogMap | 0.8689 | 0.7939 | 0.8297 |

| | | | |
|---|---|---|---|
| LogMap-Bio | 0.7443 | 0.8570 | 0.7967 |

**Table 10 Top 5 mediating ontology for small FMA SNOMED**

| | Positive hits | Composed mappings from Mediating ontology | | |
|---|---|---|---|---|
| | | PRECISION | RECALL | F-MEASURE |
| UBERON | 37 | 0.8496 | 0.3212 | 0.4661 |
| SYN | 18 | 0.8601 | 0.1637 | 0.2751 |
| MA | 12 | 0.8974 | 0.0894 | 0.1625 |
| CL | 8 | 0.8019 | 0.0458 | 0.0867 |
| BIRNLEX | 7 | 0.9003 | 0.0301 | 0.0582 |

**Table 10a Small FMA SNOMED LogMap to LogMap-Bio**

| | PRECISION | RECALL | F-MEASURE |
|---|---|---|---|
| LogMap | 0.9643 | 0.6680 | 0.7892 |
| LogMap-Bio | 0.9521 | 0.6792 | 0.7928 |

**Table 11 Top 5 mediating ontology for the whole FMA SNOMED**

| | Positive hits | Composed mappings from Mediating ontology | | |
|---|---|---|---|---|
| | | PRECISION | RECALL | F-MEASURE |
| UBERON | 35 | 0.4381 | 0.2745 | 0.3375 |
| SYN | 15 | 0.7128 | 0.1476 | 0.2446 |
| MA | 7 | 0.7685 | 0.0807 | 0.1461 |
| CL | 6 | 0.6737 | 0.0390 | 0.0737 |
| ONTOAD | 6 | 0.5303 | 0.0078 | 0.0153 |

**Table 11a Whole FMA SNOMED LogMap to LogMap-Bio**

| | PRECISION | RECALL | F-MEASURE |
|---|---|---|---|

| | | | |
|---|---|---|---|
| LogMap | 0.8753 | 0.5967 | 0.7096 |
| LogMap-Bio | 0.8220 | 0.6189 | 0.7061 |

**Table 12 Top 5 mediating ontology for small SNOMED NCI**

| | Positive hits | Composed mappings from Mediating ontology | | |
|---|---|---|---|---|
| | | PRECISION | RECALL | F-MEASURE |
| Syn | 29 | 0.8664 | 0.1332 | 0.2309 |
| CHEBI | 28 | 0.9315 | 0.1176 | 0.2088 |
| DINTO | 24 | 0.9678 | 0.0574 | 0.1084 |
| DOID | 24 | 0.7778 | 0.1287 | 0.2209 |
| NATPRO | 18 | 0.9043 | 0.0938 | 0.1699 |

**Table 12a Small SNOMED NCI LogMap to LogMap-Bio**

| | PRECISION | RECALL | F-MEASURE |
|---|---|---|---|
| LogMap | 0.8933 | 0.6641 | 0.7619 |
| LogMap-Bio | 0.8873 | 0.6719 | 0.7648 |

**Table 13 Top 5 mediating ontology for the whole SNOMED NCI**

| | Positive hits | Composed mappings from Mediating ontology | | |
|---|---|---|---|---|
| | | PRECISION | RECALL | F-MEASURE |
| CHEBI | 32 | 0.8947 | 0.1167 | 0.2065 |

| | | | | |
|---|---|---|---|---|
| DINTO | 29 | 0.9210 | 0.0699 | 0.1300 |
| DOID | 20 | 0.7106 | 0.1240 | 0.2112 |
| SYN | 19 | 0.7761 | 0.1280 | 0.2198 |
| NATPRO | 19 | 0.8740 | 0.0924 | 0.1671 |

**Table 13a Whole SNOMED NCI LogMap to LogMap-Bio**

| | PRECISION | RECALL | F-MEASURE |
|---|---|---|---|
| LogMap | 0.8677 | 0.5602 | 0.6809 |
| LogMap-Bio | 0.8590 | 0.5854 | 0.6963 |

For all the tasks in this track, SYN ontology is always in the Top 5 mediating ontologies and has a positive contribution for the F-measure. The SYN ontology uses the vocabulary in the Synapse platform (Zhang et al. 2007), This platform is used by biomedical data scientists and biological scientists. This ontology includes broad general concepts in the biological field and thus, has sematic overlap with all the three ontology. It is interesting to note that the top 5 mediating ontology for SNOMED –NCI is much different from the FMA pairs. Only SYN is the same. This difference occurs because those mediating ontologies are highly related to both SNOMED and NCI but are not very related to FMA.

For both FMA-NCI and SNOMED-NCI pairs the Top 5 mediating ontologies are same between the whole version and the small versions, and the ranking is even the same for the FMA-NCI small and whole versions. For the whole version and the small versions of the FMA SNOMED pair, the top 4 mediating ontology are same and the ranking is the same.

An examination of the hit counts between the whole and small versions shows that they are in about the same range even though the number of concepts in the whole versions are considerably larger than in the small versions. As shown in Algorithm 1 if the top 5 mediating ontologies do not change after 25 calls to BioPortal then these 5 ontologies are selected as the mediating ontologies for the source and target ontologies. This stopping condition allows efficient processing even for very large ontologies.

## 5.3 Summary

The experiments of LogMap with BioPortal as a mediating ontology provider in the anatomy track have shown that the new software component indeed improves the recall and F- measure as show in Table 4. In the Large Biomedical Track, however, only three of the six tasks show improvement in the overall performance. These are the Small FMA-SNOMED, the Small SNOMED-NCI and the whole SNOMED-NCI tasks. The problem is the large number of concepts in these ontologies result in many new mappings being found using the mediating ontologies that can significantly reduce the precision. For example, the precision of the whole FMA-NCI task and precision of the whole FMA-SNOMED are reduced. Many new mappings are included but as one can see although recall is increased, the increase cannot offset the decreased precision.

Because the mappings found using the mediating ontologies can overlap with mapping from original LogMap, the increased performance from just the mappings found by the mediating ontology are more indicative of the usefulness of this approach. For example, the precision of the new mappings in whole FMA-NCI task is 0.283 and the precision of the new mappings in small FMA-NCI is 0.4277. The recall of these two tasks is also very low, which is 0.0225 and 0.07266. This result shows that the use of the mediating ontologies for the FMA-NCI pairs did not improve the overall LogMap final recall but damaged the precision by adding some incorrect mappings. These low precisions from the mappings produced by the mediating ontologies caused the F-Measure to be lowered.

# 6 Other OA systems and BIOPORTAL

To demonstrate the use of BioPortal as a source for mediating ontologies with other OA systems, an interface was developed for those systems that participated in the 2012, 2013, and 2014 OAEI competitions. Most OA systems could complete the OAEI anatomy track; however, some were not able to perform in the large biomedical track. For example, some of them threw errors, and some produced an empty alignment file. For this reason, only the anatomy track was used to examine how the use of BioPortal affected the performance of these other OA systems.

Of the OA systems participating in the anatomy track, several were over optimized for the anatomy track competition and could not make alignments to the mediating ontologies in order to produce composed mappings. Because of their limitations, they could not use BioPortal as a source for meditating ontologies. The result is that only 7 OA systems could be used to evaluate the software to use BioPortal for its mediating ontologies.

In table 14, the rows with "-Bio" added the name of OA systems indicate that BioPortal was used as a provider of mediating ontologies. The use of these mediating ontologies follows the parameters and the top five mediating ontologies as described in section 5. 1.

By using BioPortal, the recall of each OA system is increased though in some cases very slightly. This increase in recall indicates that the use of BioPortal helped find additional correct mappings. The F-measure, however, is only increased for AgreementMaker Light (AML) (Faria et al. 2013) and Hertuda (Hertling 2012).
Because the input source and target ontologies are the same as in section 5, the top 5 mediating ontologies are the same as in the section 5. The composed mappings for each OA system, however, is different because they are determined by mapping from the source concept to the mediating ontology and the target concept to the mediating ontology and then seeing if the mapping results in the identical concept in the mediating ontology.

The performance of composed mappings for each OA system is dependent on both its ability to produce high recall and precision for the alignment algorithm it uses. The OA system must perform its own mapping from the source to the mediating ontology and from the target to the mediating ontology. If the recall performance of the origin OA system is low, then the OA system will have difficulty finding correct mappings to the mediating ontology from both the source and target ontologies; therefore, producing correct composed mappings between the source and target ontologies will be difficult. The performance of using mediating ontology is therefore limited based on the alignment capabilities of the original OA system.

**Table 14 Other OA systems performance using BioPortal**

| | PRECISION | RECALL | F-MEASURE |
|---|---|---|---|
| AML | 0.9548 | 0.8219 | 0.8834 |
| AML-Bio | 0.9319 | 0.8938 | **0.9125** |
| AOTL | 0.7024 | 0.0778 | 0.1401 |
| AOTL -Bio | 0.6296 | 0.0785 | 0.1396 |
| GOMMA | 0.9505 | 0.7975 | 0.8673 |
| GOMMA -Bio | 0.9395 | 0.7995 | 0.8639 |
| Hertuda | 0.6892 | 0.6728 | 0.6809 |
| Hertuda -Bio | 0.6889 | 0.6735 | **0.6811** |
| AOT | 0.4352 | 0.7751 | 0.5574 |
| AOT-Bio | 0.4325 | 0.7777 | 0.5559 |
| StringAuto | 0.8947 | 0.7790 | 0.8329 |
| StringAuto -Bio | **0.7925** | 0.7909 | 0.7917 |
| Wiki 2013 | 0.9802 | 0.6517 | 0.7829 |
| Wiki 2013 -Bio | 0.9661 | 0.6570 | 0.7821 |

In order to examine how the precision of the OA system affects the composed mapping, the performance measures for just the new mapping produced using the mediating ontologies were calculated and are provided in Table 15.

**Table 15 Performance new mapping from composed mappings from the mediating ontologies**

| New Mapping performace | PRECISION | RECALL | F-MEASURE |
|---|---|---|---|
| AML | 0.7006 | 0.0726 | 0.1315 |
| AOTL | 0.0476 | 0.0007 | 0.0013 |
| GOMMA | 0.1667 | 0.0020 | 0.0039 |
| Hertuda | 0.0010 | 0.0007 | 0.0008 |
| AOT | 0.1538 | 0.0026 | 0.0052 |
| StringAuto | 0.0914 | 0.0119 | 0.0210 |
| Wiki 2013 | 0.3478 | 0.0053 | 0.0104 |

In Figure 4, the precision from only the new mappings is compared to the precision of the original OA system. Except for AOT system, the precision of the new mapping and the precision of the original OA system are related. The ability to find correct composed mappings by using mediating ontologies is dependent on the precision of the original OA system. Since several OA systems have poor precision, not enough OA systems are available to determine unquestionably that the use of mediating ontologies in BioPortal has a positive effect on the alignment results.

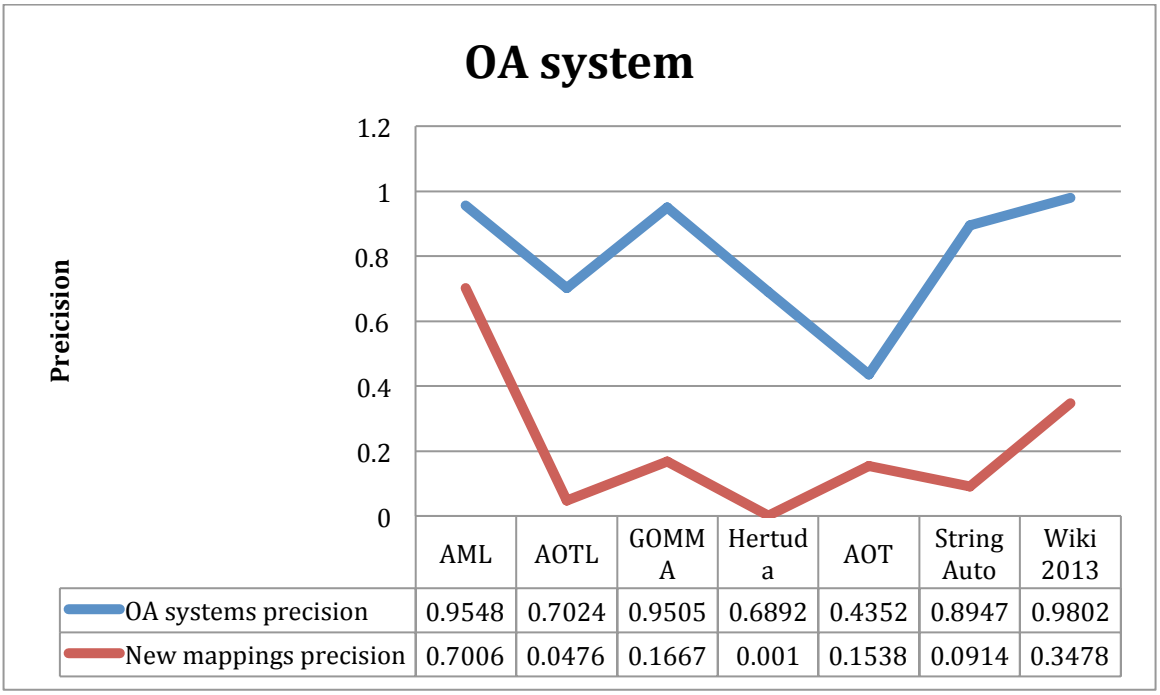


**Figure 4 Precision comparison between original OA system and that of new composed mappings only**

To summarize, the evaluation showed that if the OA system has low precision, then the precision for the composed mappings is low. Even though the mediating ontology technique improves recall, the original recall performance of the OA system still needs to be reasonable such as for AOT.

If more OA systems become available that are not overly optimized then more experiments may be undertaken to determine how the recall and precision of the original OA system can affect the performance of using mediating ontologies.

# 7 CONCLUSIONS AND FUTURE WORK

This thesis work exploits BioPortal as a source of mediating ontologies rather than pre-selecting a mediating ontology for each alignment task. Experiments have been done using LogMap as the base OA system and BioPortal as the mediating ontology provider for both the Anatomy track and the Large Biomedical Ontology Track of the Ontology Alignment Evaluation Initiative (OAEI). At the beginning of the research, the goal was to

find the best ontology from BioPortal to be used as a mediating ontology. But further experiments showed that using one mediating ontology that is most similar to the target and source ontologies based on the number of positive query responses may result in producing a set of mapping that have a large overlap with the set of mappings produced by the original OA system without using BioPortal. Based on that discovery, the approach to use the top 5 mediating ontology was developed. Further experimentation lead to the method of parameterizing the number of mediating ontologies with the confidence level of the mapping as a way to increase recall without decreasing precision as much.

Two types of evaluation of the performance of using mediating ontologies provide by BioPortal have been done. First experiments with a leading OA system LogMap were performed with the new component to use BioPortal for both the OAEI anatomy track and large biomedical ontologies track. Then the new software component was used with other available OA systems that participated in the OAEI in anatomy track in order to examine how effective it is with other such systems.

This thesis research makes the following contributions:

1) BioPortal has never been used as mediating ontology provider. This thesis is the first to provide a software component that can be used by OA systems to use BioPortal in such a role.
2) In the bioinformatics domain, an efficient approach to select mediating ontologies rather than having a pre-selected ontology for alignment tasks is presented.
3) For two leading OA systems LogMap and AML, the software using BioPortal increased both the recall and F-measure for the anatomy track. For LogMap in the large biomedical ontologies track, the new feature increased the F-measure and recall for FMA-SNOMED small fragment task, SNONMED NCI whole ontologies task and SNONMED NCI small fragments task.
4) The evaluation of the software using BioPortal with other OA systems revealed that many of them are over specialized for the anatomy track. This specialization

makes some OA systems unable to be used to align other ontologies outside of the anatomy track and, therefore, the use of mediating ontologies is not possible.

5) The evaluation of the software using BioPortal with other available OA systems provides a better understanding of the limitation of mediating ontologies in the OA process. The performance calculated on the mappings found by using mediating ontologies is dependent on the performance of origin OA system since finding the composed mappings depends on OA systems ability to finding mappings to the mediating ontologies.

Future research for the use of mediating ontologies in the OA process includes:

1) The number of OA system is too limited. Most of the OA systems that participate in OAEI are over specialized. If more OA systems become available, the use of mediating ontologies can be more fully explored.
2) The software developed using BioPortal as the mediating ontology provider could be adapted to other domains than just biomedical and bioinformatics fields. Its use with other ontology repositories should be investigated.

# REFERENCES


Addagada, Sridevi. 2007. “Indexing and Searching Document Collections Using Lucene.” University of New Orleans.

Cruz, I. F., Stroe C., Caimi F., Fabiani A., Pesquita C., Couto F., & Palmonari M. 2011. “Using AgreementMaker to Align Ontologies for OAEI 2011.” *Proceedings of the Sixth International Workshop on Ontology Matching*, 114–21. http://disi.unitn.it/~p2p/OM-2011/om2011_proceedings.pdf#page=124.

Dragisic, Zlatan, Kai Eckert, Alfio Ferrara, Roger Granada, Valentina Ivanova, Ernesto Jim, Dominique Ritze, Pavel Shvaiko, and Alessandro Solimando. 2014. “Results of the Ontology Alignment Evaluation Initiative 2014.”

Euzenat, Jérôme. 2007. “Semantic Precision and Recall for Ontology Alignment Evaluation.” *IJCAI International Joint Conference on Artificial Intelligence*, 348–53.

Euzenat, Jérôme, and Pavel Shvaiko. 2007. *Ontology Matching*. *Booksgooglecom*. doi:10.1007/978-3-540-49612-0.

Faria, Daniel, Catia Pesquita, Emanuel Santos, Matteo Palmonari, Isabel F. Cruz, and Francisco M. Couto. 2013. “The AgreementMakerLight Ontology Matching System.” *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8185 LNCS: 527–41. doi:10.1007/978-3-642-41030-7_38.

Grau, Bernardo Cuenca, Zlatan Dragisic, Kai Eckert, Alfio Ferrara, Roger Granada, Valentina Ivanova, Ernesto Jim, and Dominique Ritze. 2013. “Results of the Ontology Alignment Evaluation Initiative 2013.”

Gross, Anika, Michael Hartung, Toralf Kirsten, and Erhard Rahm. 2011. “Mapping Composition for Matching Large Life Science Ontologies.” In *2nd International Conference on Biomedical Ontology*. http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.225.7501.

Gruber, TR. 1993. “A Translation Approach to Portable Ontology Specifications.” *Knowledge Acquisition* 5 (2): 199–220. http://secs.ceas.uc.edu/~mazlack/ECE.716.Sp2011/Semantic.Web.Ontology.Papers/Gruber.93a.pdf.

Hertling, Sven. 2012. “Hertuda Results for OEAI 2012.” *ISWC Workshop*.

Hertling, Sven, and Heiko Paulheim. 2012. “WikiMatch–Using Wikipedia for Ontology

Matching.” *Disi.unitn.it*. http://disi.unitn.it/~p2p/OM-2012/om2012_Tpaper4.pdf.

Jiménez-Ruiz, E, BC Grau, Yujiao Zhou, and Ian Horrocks. 2012. “Large-Scale Interactive Ontology Matching: Algorithms and Implementation.” *ECAI*, no. ii: 0–5. http://ebooks.iospress.nl/Download/Pdf/7013.

José-Luis Aguirre, Kai Eckert, Jérôme Euzenat, Alfio Ferrara, Willem Robert van Hage, Laura Hollink, Christian Meilicke, Andriy Nikolov, Dominique Ritze, François Scharffe, Pavel Shvaiko, Ondrej Sváb-Zamazal, Cássia Trojahn dos Santos, Ernesto Jiménez-Rui, Benjamin Zapilko. 2012. “Results of the Ontology Alignment Evaluation Initiative 2012.” In *Ontology Matching*. http://hal.inria.fr/hal-00768409/.

Kamps, C, and S Safe. 1987. “Binding of Polynuclear Aromatic Hydrocarbons to the Rat 4S Cytosolic Binding Protein: Structure-Activity Relationships.” *Cancer Letters* 34 (2): 129–37. doi:10.1016/0304-3835(87)90003-6.

Kensche, David, Christoph Quix, Xiang Li, and Yong Li. 2007. “GeRoMeSuite: A System for Holistic Generic Model Management.” In *33rd International Conference on Very Large Data Bases*. http://dl.acm.org/citation.cfm?id=1326004.

Li Ding, Rong Pan, Tim Finin, Anupam Joshi, Yun Peng, and Pranam Kolari. 2005. “Finding and Ranking Knowledge on the Semantic Web.” In *Proceedings of the 4th International Semantic Web Conference*. http://link.springer.com/chapter/10.1007/11574620_14.

Mungall, Christopher J, Carlo Torniai, Georgios V Gkoutos, Suzanna E Lewis, and Melissa a Haendel. 2012. “Uberon, an Integrative Multi-Species Anatomy Ontology.” *Genome Biology* 13 (1). BioMed Central Ltd: R5. doi:10.1186/gb-2012-13-1-r5.

Nakayama, Kotaro, Takahiro Hara, and Shojiro Nishio. 2008. “Wikipedia Link Structure and Text Mining for Semantic Relation Extraction towards a Huge Scale Global Web Ontology.” *CEUR Workshop Proceedings* 334: 59–73.

Noy, Natalya F, Nigam H Shah, Patricia L Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, et al. 2009. “BioPortal: Ontologies and Integrated Data Resources at the Click of a Mouse.” *Nucleic Acids Research* 37 (Web Server issue): W170–73. doi:10.1093/nar/gkp440.

Prateek Jain, Peter Z. Yeh, Kunal Verma, Reymonrod G. Vasquez, Mariana Damova, Pascal Hitzler, Amit P. Sheth. 2011. *Contextual Ontology Alignment of Lod with an Upper Ontology: A Case Study with Proton*. *The Semantic Web: Research and Applications*. http://www.springerlink.com/index/530M84420Q6G8537.pdf.

Quix, Christoph, Pratanu Roy, and David Kensche. 2011. “Automatic Selection of

Background Knowledge for Ontology Matching." *Proceedings of the International Workshop on Semantic Web Information Management - SWIM '11* 5. New York, New York, USA: ACM Press: 1–7. doi:10.1145/1999299.1999304.

Reynaud, Chantal, and Brigitte Safar. 2007. "Exploiting WordNet as Background Knowledge." *CEUR Workshop Proceedings* 304.

Rubin, DL, DA Moreira, Pradip P. Kanjamala, and Mark A. Musen. 2008. "BioPortal: A Web Portal to Biomedical Ontologies." In *AAAI Spring Symposium Series, Symbiotic Relationships between Semantic Web and Knowledge Engineering*, 1–4. http://www.aaai.org/Papers/Symposia/Spring/2008/SS-08-07/SS08-07-011.pdf.

Shvaiko, P., and J. Euzenat. 2013. "Ontology Matching: State of the Art and Future Challenges." *IEEE Transactions on Knowledge and Data Engineering* 25 (1): 158–76. doi:10.1109/TKDE.2011.253.

Silwal, P. 2012. "ONTOLOGY ALIGNMENT USING SEMANTIC SIMILARITY WITH REFERENCE ONTOLOGIES." *Vasa*. Miami University. http://medcontent.metapress.com/index/A65RM03P4874243N.pdf.

Taye, Mm, and Na Alalwan. 2010. "Ontology Alignment Technique for Improving Semantic Integration." In *The Fourth International Conference on Advances in Semantic Processing*, 13–18. http://www.thinkmind.org/index.php?view=article&articleid=semapro_2010_1_30_50049.

Zhang, Wuxue, Yong Zhang, Hui Zheng, Chen Zhang, Wei Xiong, John G Olyarchuk, Michael Walker, et al. 2007. "SynDB: A Synapse Protein DataBase Based on Synapse Ontology." *Nucleic Acids Research* 35 (Database issue): D737–41. doi:10.1093/nar/gkl876.